Qeios

Edit Away and My Face Will not Stay: Personal Biometric Defense against Malicious Generative Editing

Hanhui Wang[1,*], Yihua Zhang[2,*], Ruizheng Bai[3], Yue Zhao[1], Sijia Liu[2], Zhengzhong Tu[3]

[1]University of Southern California [2]Michigan State University [3]Texas A&M University

hanhuiwa@usc.edu, zhan1908@msu.edu, tzz@tamu.edu

*Equal contribution

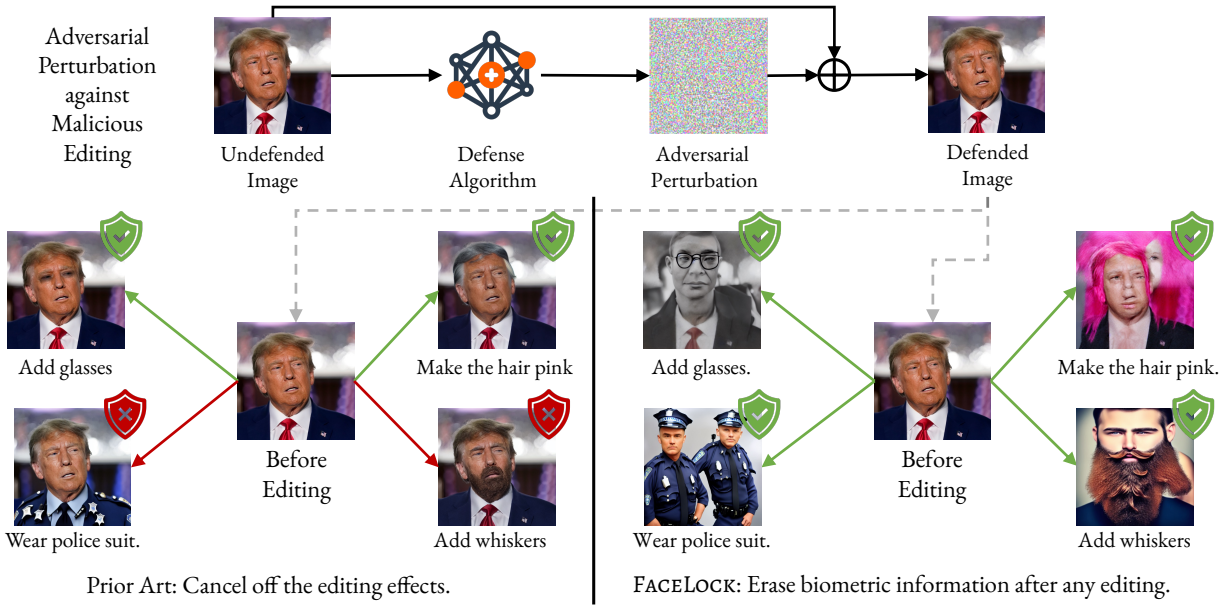

Figure 1. An illustration of adversarial perturbation generation for safeguarding personal images from malicious editing. Perturbations generated by prior work [1, 2] aim to cancel off editing effects, resulting in instability due to the diversity of editing instructions. In contrast, FaceLock does not prevent edits from being applied but instead erases critical biometric information (*e.g.*, human facial features) after editing, making it agnostic to specific prompts and achieving superior performance.

Abstract

Recent advancements in diffusion models have made generative image editing more accessible than ever. While these developments allow users to generate creative edits with ease, they also raise significant ethical concerns, particularly regarding malicious edits to human portraits that threaten individuals' privacy and identity security. Existing general-purpose image protection methods primarily focus on generating adversarial perturbations to nullify edit effects. However, these approaches often exhibit instability to protect against diverse editing requests. In this work, we introduce a novel perspective to personal human portrait protection against malicious editing. Unlike traditional methods aiming to prevent edits from taking effect, our method, FaceLock*, optimizes adversarial perturbations to ensure that original biometric information—such as facial features—is either destroyed or substantially altered post-editing, rendering the subject in the edited output biometrically unrecognizable. Our approach innovatively integrates facial recognition and visual perception factors into the perturbation optimization process, ensuring robust protection against a variety of editing attempts. Besides, we shed light on several critical issues with commonly used evaluation metrics in image editing and reveal cheating methods by which they can be easily manipulated, leading to deceptive assessments of protection. Through extensive experiments, we demonstrate that* FaceLock *significantly outperforms all baselines in defense performance against a wide range of malicious edits. Moreover, our method also exhibits strong robustness against purifi-*

Preprint — https://doi.org/10.32388/NPNABI

cation techniques. Comprehensive ablation studies confirm the stability and broad applicability of our method across diverse diffusion-based editing algorithms. Our work not only advances the state-of-the-art in biometric defense but also sets the foundation for more secure and privacy-preserving practices in image editing. The code is publicly available at: https://github.com/taco-group/FaceLock.

1. Introduction

Image editing has advanced at an unprecedented rate due to the rise of diffusion-based techniques, making it possible to produce edits that are indistinguishable from reality [3–14]. This rapid development has led to tools capable of seamlessly modifying visual content, with edits so convincing that they are often impossible to differentiate from the original image. While this progress opens up creative possibilities, it also brings significant ethical and societal challenges.

The power of these editing techniques has led to severe ethical implications [15–17]. Recent incidents, such as the widely discussed manipulation of Taylor Swift's images [18] and the proliferation of pornographic content affecting Korean schools [19], underscore the urgent need to address the risks associated with malicious image editing. These incidents have highlighted growing concerns about how personal images, particularly those depicting individuals' faces, can be misused once they are posted online [13, 14, 20]. Protecting such images from unauthorized and malicious edits has thus become an important topic of research [21–23].

To address this challenge, several recent attempts [1, 2, 24–28] have focused on using adversarial perturbations, which are imperceptible to human eyes but are intended to negate the effects of editing when such images are used as inputs to diffusion-based editing algorithms. These perturbations aim to protect personal images by preventing the success of the intended edits (see **Fig. 1** for an illustration). However, current methods suffer from instability [1, 2, 27, 28] and simple purification methods. Specifically, while they are effective for certain types of editing instructions, they fail against others, largely due to the inherent diversity and versatility of editing prompts. The underlying issue is that as long as existing methods continue to focus on 'canceling off editing effects', the inconsistency of results is inevitable. The diversity in editing prompts and the complexity of generative diffusion models make it difficult for such approaches to generalize effectively.

The rationale behind current defense methods is to ensure that the edited image does not meet the requirements of a successful image editing task. To understand this more deeply, we first revisit what constitutes a successful image editing task: it should accurately reflect the editing instruction while preserving the original, irrelevant visual features, such as those related to the subject's identity, including facial features. The latter requirement, which has been largely overlooked, provides an opportunity for a new defense strategy. Instead of attempting to cancel out edits, here, we ask:

> ***(Q)*** *Can we design adversarial perturbations that cause edited images to lose their biometric information, making the edited image biometrically unrecognizable and thereby causing the edit to fail?*

Through a series of algorithmic designs, we demonstrate that creating adversarial perturbations that disrupt facial recognition while also introducing distinct visual disparities in facial features is far from trivial. To address this, we propose FACELOCK, which strategically integrates a state-of-the-art facial recognition model into the diffusion loop as an adversary while also penalizing feature embeddings to achieve visual dissimilarity. By doing so, our method not only disrupts facial recognition but also ensures significant visual differences from the original, providing robust protection against malicious editing, see **Fig. 1** for a comparison between FACELOCK and prior arts. To this end, we summarize our contributions as follows:

- We present a novel perspective for protecting personal images from malicious editing, focusing on making biometric features unrecognizable after edits.
- We develop a new algorithm, FACELOCK, that incorporates facial recognition models and feature embedding penalties to effectively protect against diffusion-based image editing.
- We conduct a critical analysis of the quantitative evaluation metrics commonly used in image editing tasks, exposing their vulnerabilities and highlighting the potential for manipulation to achieve deceptive results.
- Through extensive experiments, we demonstrate that FACELOCK effectively alters human facial features against various editing prompts, achieving superior defense performance compared to baselines. We also show that FACELOCK generalizes well to multiple diffusion-based algorithms and exhibits inherent robustness against purification methods.

2. Related Work

Generative editing models. Recent advances in latent diffusion models [29] have demonstrated superior image editing capabilities through instructions and prompt editing [3–6]. Most recent methods [7, 8] combine diffusion models with large language models for understanding text prompts. InstructPix2Pix [9] leverages a fine-tuned version of GPT-3 and images generated from SD and achieves on-the-fly image editing without further per-sample finetuning. On the other hand, many such models also allow personalized image editing [10, 11]. DreamBooth [12] learns a unique identifier and class type of an object by finetuning a pretrained text-to-image model with a few images. SwapAnything [14] and Photoswap [13] allow for personal content editing by swapping faces and objects between two images. In the gen-

erative era, these tools offer unprecedented creative freedom but also raise ethical questions on privacy and malicious image editing, which motivate us to conduct this work.

Defense against malicious editing. Adversarial samples are clean samples manipulated intentionally to fool a machine learning model, often done by perturbing the image with an imperceptible small noise. Under a white-box setting, gradient-based methods, such as fast gradient sign method (FGSM), projected gradient decent (PGD) [30] and Carlini & Wagner (CW) attack [31], are among the most effective techniques in generating adversarial examples in classification models. Recent works like PhotoGuard, Editshield, AdvDM [1, 2, 27] have extended gradient based methods to diffusion models and aim to protect images from malicious editing. PhotoGuard demonstrated an effective encoder attack mechanism by perturbing the source image towards an unrelated target image, e.g. an image of gray background. In particular, let $\mathcal{E}$ be the encoder, $\mathbf{z}_{\text{target}}$ be the latent representation of the target image. Under a attack budget ϵ, PhotoGuard aims to optimize:

$$\delta_{\text{Encoder}} = \arg\min_{\|\delta\|_\infty \leq \epsilon} \|\mathcal{E}(\mathbf{x}+\delta) - \mathbf{z}_{\text{target}}\|. \quad (1)$$

Yet, the protection can be less effective if the image is slightly transformed. PhotoGuard takes a step further by considering expectations over transformation:

$$\max_{\mathbf{x_p}} \mathbb{E}_{f\sim\mathcal{F}} \left[\text{Dist}(\mathcal{E}(f(\mathbf{x_p})), \mathcal{E}(\mathbf{x}))\right] - \beta \cdot \|\mathbf{x_p} - \mathbf{x}\|_2^2, \quad (2)$$

where $\mathbf{x}$ is the source image, $\mathbf{x_p}$ is the perturbed image, and $\mathcal{F}$ is a distribution over a set of transformations.

However, these approaches are typically less robust, as the gradients are highly dependent on model architecture and parameters. Distraction Is All you Need [25] circumvent this by attacking the cross attention mechanism between image and editing instruction, so diffusion models misinterpret the target editing regions. Glaze and Nightshade [24, 26] instead perturb the image towards a completely different image with another style or concept. These approaches make the image less susceptible to the specificities of model architecture and is generally more robust across different models.

The adversarial techniques mentioned above primarily protect portrait images by interfering with the editing process. However, nullifying the editing process does not always safeguard facial features or biometric information. We propose a novel way of protecting images by incorporating facial recognition model into the perturbation process. Although the model is capable of editing the image according to the prompt, we ensure that the facial features are altered or destroyed during the process.

Facial recognition. Recent facial recognition works [32–38] have proposed several margin-based softmax loss functions to enhance the discriminative power and feature extraction ability of facial recognition models. CVLFace [32] utilizes these models to extract features from two images and computes the cosine similarity between these features to verify a person's identity. In addition, recent works [39–43] also leverage synthetic images during training for enhanced privacy protection, highlighting the need for privacy protection in image editing as well. Our approach builds on top of CVLFace and protects biometric information by minimizing the cosine similarity between features. Our work is the first application of facial recognition on perturbation generation, enabling a new axis of identity protection.

3. FaceLock: Adversarial Perturbations for Biometrics Erasure

Figure 2. Illustration of the two requirements of image editing task: prompt fidelity and image integrity. (a) Source image before editing; (b) A successful editing example holding both metrics; (c) Failure case due to the lack of prompt fidelity leading to under-editing and (d) the lack of the image integrity leading to over-editing.

What defines a successful image editing task? Before introducing our proposed method for safeguarding human portrait images from malicious edits, we revisit the criteria for a successful image editing outcome. Specifically, we propose that a successful text-guided image editing hinges on two critical requirements: ❶ **prompt fidelity**, and ❷ **image integrity**. **Prompt fidelity** requires that the edit accurately reflects the instructions provided in the prompt. For instance, as shown in **Fig. 2**, a successful edit replaces the person's clothing with a police uniform as instructed by the prompt. Meanwhile, **image integrity** requires that other elements in the image remain intact after editing. Although this requirement is less explicit than prompt fidelity, it defines the essence of image editing and differentiates it from general text-to-image generation tasks. As illustrated in Fig. 2, aside from the change in attire, the edited image should retain as much of the subject's original appearance as possible, including facial features, poses, and other details. While prompt fidelity has been emphasized and extensively studied [7, 9, 12, 29], image integrity remains long-overlooked and underexplored in literature. Next, we will demonstrate how this holistic view of image editing can provide new insights into protecting human portraits from malicious edits.

A new direction for defending against malicious editing. As discussed above, to safeguard personal images from malicious editing, the defender must ensure that at least one of the two requirements is not met. Previous works have

FR=1.0	FR=0.972	FR=0.901	FR=0.273	FR=0.658	FR=0.093
	✗ Visual Change	✗ Visual Change	✗ Visual Change	✗ Visual Change	✓ Visual Change
(a) Source Image	(b) No Protection	(c) Design I: CVL	(d) Design II: CVL-D	(e) Design III: CVL-D + Pixel	(f) FACELOCK

Figure 3. Source and edited images generated from different protection methods based on the instruction "*Let the person wear a police suit*". The FR score below each image represents the facial representation similarity between the edited and source images and scores marked in red indicate insignificant changes biometric recognition results by CVLFACE compared to source image. 'CVL' refers to perturbations generated targeting the CVLFACE model alone. 'CVL-D' represents protection targeting both the CVLFACE model and the diffusion model, while 'CVL-DP' incorporates an auxiliary loss to enforce pixel-level disparity between the edited and source images. FACELOCK targets the CVLFACE and diffusion model, aiming to enhance the disparity between the feature embeddings of the decoded and original images.

primarily focused on generating adversarial perturbations to prevent edits from taking effect, thereby reducing prompt fidelity [1, 2, 24, 26, 27]. However, these approaches often suffer from instability and are effective only for a limited range of editing instructions, resulting in poor generalization. The core issue is the versatility of editing instructions—making it unlikely that a single perturbation can defend against all potential prompts. Therefore, we explore a new direction: optimizing perturbations to destroy biometric information after editing, rendering the edited image biometrically unrecognizable and thereby causing the edit to fail.

Adversarial perturbation for facial disruption is nontrivial. The goal of our defense method is to disrupt human facial features during the sampling process in diffusion-based editing models. **Design I (CVL): Perturbation against facial recognition models.** A straightforward approach is to apply an adversarial perturbation against a state-of-the-art (SOTA) facial recognition model, such as the CVLFACE model [32], and use the perturbed image as input to the image editing model. However, as shown in **Fig. 3(c)**, the perturbation that successfully fools the CVLFACE model does not persist through the diffusion model's sampling process, resulting in an edited image with minimal disruption to facial features, as indicated by both the high facial similarity (FR) score and the visually similar appearance. The underlying issue with this approach is that the perturbations are generated independently of the diffusion process. Prior work [44] highlights that diffusion models possess an inherent ability to "purify" adversarial perturbations through their sampling process.

Design II (CVL-D): Perturbation against diffusion with CVLFACE model in the loop. To address this, we incorporate the CVLFACE model into the diffusion process and design a method to directly interfere with the sampling stage. Given the high computational costs of disrupting the entire diffusion process, we instead bypass this step once the latent representation of the input image is obtained. The perturbation is then optimized by a facial recognition loss that maximizes the biometric disparity between the decoded image and the source input:

$$\boldsymbol{\delta} = \underset{\|\boldsymbol{\delta}\|_\infty \leq \epsilon}{\arg\max}\, f_{\text{FR}}(\mathcal{D}(\mathcal{E}(\mathbf{x}+\boldsymbol{\delta})), \mathbf{x}), \tag{3}$$

where $\mathcal{D}$ and $\mathcal{E}$ denote the decoder and encoder used by the diffusion model, respectively, and $f_{\text{FR}}(\cdot,\cdot)$ computes the facial similarity score. As shown in **Fig. 3(d)**, while this method significantly reduces the facial recognition similarity score, the edited image still resembles the original subject, suggesting room for further improvement in visual effects.

Design III (CVL-DP): Perturbation against diffusion facial similarity with pixel-level penalty. To enhance the visual disparity between the edited image and the source image, we introduce a pixel-level loss focused on facial regions defined by a mask:

$$\boldsymbol{\delta} = \underset{|\boldsymbol{\delta}|_\infty \leq \epsilon}{\arg\max}\, f_{\text{FR}}(\mathcal{D}(\mathcal{E}(\mathbf{x}+\boldsymbol{\delta})), \mathbf{x}) + \lambda \|\boldsymbol{\delta} \odot \mathbf{m}\|_2, \tag{4}$$

where $\mathbf{m}$ defines the facial region extracted by the CVLFACE model. However, as shown in **Fig. 3(e)**, the pixel-level loss primarily results in color shifts rather than significant distortion of the subject's facial features. This limitation motivated the development of FACELOCK, which aims to generate perturbations that enhance both facial dissimilarity scores and visual facial discrepancies.

FACELOCK: Perturbation optimization on facial disruption and feature embedding disparity. The lesson from CVL-DP indicates that pixel-level changes do not necessarily lead to distinct visual facial features. Thus, we transition to a more effective feature-level approach, using pretrained convolutional neural networks to extract and maximize the difference between high-level feature embeddings of the decoded and source image:

$$\boldsymbol{\delta} = \underset{|\boldsymbol{\delta}|_\infty \leq \epsilon}{\arg\max}\, f_{\text{FR}}(\mathcal{D}(\mathcal{E}(\mathbf{x}+\boldsymbol{\delta})), \mathbf{x}) + \lambda f_{\text{FE}}(\mathcal{D}(\mathcal{E}(\mathbf{x}+\boldsymbol{\delta}), \mathbf{x}), \tag{5}$$

where $f_{\text{FE}}(\cdot,\cdot)$ extracts feature embeddings from the input images and compute the distance between them. To solve (5), the widely used projected gradient descent (PGD) [30]

method can be employed. We refer more implementation details in Sec. 5.

4. Pitfalls in The Widely-Used Quantitative Evaluation Metrics for Image Editing Tasks

In this section, we begin by providing a critical analysis of existing quantitative evaluation metrics for image editing tasks [45–47]. For the first time, we highlight potential pitfalls in these widely accepted metrics, particularly how they can be easily manipulated to achieve deceptively high scores. Finally, we introduce two new, more robust metrics for evaluating human portrait editing. Detailed mathematical descriptions of the quantitative evaluation metrics discussed in this section can be found in Appendix A.

Existing quantitative metrics suffer from pitfalls and can be manipulated for misleading performance. As discussed in §3, the evaluation of general image editing tasks should consider two aspects: prompt fidelity and image integrity. However, all existing quantitative metrics, including CLIP scores [46], SSIM, and PSNR primarily focus on the former, namely how well the editing instruction is reflected in the edited image. In the following, we revisit each of these metrics and demonstrate the intrinsic pitfalls in their design.

CLIP-based scores overemphasize the presence of elements from the editing instructions, often prioritizing over-editing. CLIP-based scores are widely used to assess prompt fidelity by measuring the cosine similarity between the CLIP text embedding of the editing prompt and the visual embedding difference between the edited and source images. While this metric effectively indicates whether the edit has taken effect, it tends to overemphasize the presence of specific elements in the edited image. **Fig. 4** shows a contradictory CLIP score ranking compared to the visual editing quality. Although Fig. 4(b) demonstrates a visually balanced outcome between the editing effect 'turn the hair pink' and preserving other irrelevant (especially facial) features, the CLIP-based score still assigns higher values to Fig. 4(c) and (d) simply because they show stronger 'pink hair' effects, even if the subject's identity has been completely altered. Therefore, CLIP-based scores can easily prioritize over-editing and be manipulated by replicating elements from the editing instructions.

SSIM and PSNR over-rely on differences between the edited image and the undefended source, potentially leading to a false sense of successful defense. Unlike CLIP-based scores, metrics such as SSIM and PSNR evaluate whether a defense against editing is successful by comparing the pixel-level statistical differences between the edited images with and without defense. While comparing against the edited image without defense can be effective in some scenarios, concluding that a defense is successful simply because the defended image differs from the undefended one is premature. For example, in **Fig. 5**, Fig. 5(b) demonstrates

Figure 4. CLIP score (CLIP-S) of different editing results. The CLIP score provides a contradictory ranking (III > II > I) compared to the visual quality (I > II > III), as it overemphasizes the presence of elements from the editing prompt, thereby favoring over-editing.

a successful edit based on the instruction 'Let the person wear a hat.' While Fig. 5(c) shows a genuinely successful defense, Fig. 5(d) is incorrectly assigned a lower SSIM/PSNR score (where lower scores indicate better defense). This suggests a greater pixel-level statistical distance from Fig. 5(b) compared to Fig. 5(c). However, this assessment is flawed, as Fig. 5(d) clearly represents a failed defense, given that a green hat has been applied to the source image. The pixel statistics-based score is misleading simply because the color of the hat differs from that in Fig. 5(b). Therefore, treating the edited image without defense as a gold standard is risky, as the variability of editing effects, even with a single instruction, must be considered.

Figure 5. SSIM and PSNR scores of different defense methods. Although Defense I (b) demonstrates a successful defense, Defense II (d) is assigned a much lower (better) SSIM and PSNR score simply due to its larger pixel-level statistical difference from (b). SSIM and PSNR treat the edited image *w/o* defense as the gold standard, without accounting for the diversity of possible editing outcomes, which can lead to a false sense of defense success.

LPIPS score as a more robust metric for prompt fidelity evaluation. To address the limitations of pixel-level statistics used by SSIM and PSNR, we propose using the Learned Perceptual Image Patch Similarity (LPIPS [48]) score to evaluate the similarity between edited images. Unlike traditional similarity metrics, LPIPS leverages pretrained neural networks to quantify perceptual differences by comparing high-level semantic features of images, offering a more robust assessment of protection effectiveness. We believe this approach can help mitigate the generalization issues associated with relying on a single reference image, as highlighted in the analysis above.

Table 1. Quantitative evaluation on prompt fidelity (CLIP-S, PSNR, SSIM, LPIPS) and image integrity (CLIP-I, FR). Arrows (↑ or ↓) indicate whether a higher or lower value is preferred for a successful defense. All results are averaged over 5 different random seeds for editing. Results in the form $a \pm b$ represent mean a with std b. The best result within each evaluation metric is highlighted in **bold**.

	Prompt Fidelity				Image Integrity	
Method	**CLIP-S ↓**	**PSNR ↓**	**SSIM ↓**	**LPIPS ↑**	**CLIP-I ↓**	**FR ↓**
No Defense	0.118±0.037	-	-	-	0.808±0.074	0.833±0.111
PhotoGuard Encoder attack	**0.108**±0.030	**15.44**±2.01	0.612±0.056	0.403±0.071	0.670±0.118	0.590±0.264
EditShield	0.110±0.026	17.74±2.20	0.593±0.072	0.382±0.071	0.677±0.096	0.641±0.231
Untargeted Encoder attack	0.116±0.023	16.74±2.27	**0.589**±0.084	0.371±0.094	0.653±0.090	0.563±0.236
CW L2 attack	0.115±0.031	19.64±2.46	0.701±0.060	0.247±0.062	0.733±0.089	0.725±0.173
VAE attack	0.114±0.034	19.40±1.70	0.715±0.039	0.251±0.060	0.786±0.061	0.846±0.097
FACELOCK (ours)	0.114±0.024	17.11±2.36	**0.589**±0.079	**0.436**±0.065	**0.648**±0.089	**0.315**±0.109

Facial recognition similarity score for image integrity evaluation. In this work, we propose evaluating image integrity as a means of assessing defense performance. However, we acknowledge that developing a cost-effective metric for general image integrity—defined as retaining all elements irrelevant to the editing—is challenging due to the diversity of elements present in an image. Therefore, we focus specifically on how well *human facial details* are preserved after editing, under the assumption that facial features are not altered. For this purpose, we use the facial recognition (**FR**) similarity score to compare the subjects in the edited and source images. Generally, if the edited image does not statistically (in terms of FR score) and visually resemble the original subject, it indicates a successful defense. Additionally, we use the cosine similarity of the CLIP score (CLIP-I) between the edited and source images as a reference indicator on the general preservation effect.

5. Experiments

5.1. Experiment Setup

Models and dataset. We adopt the widely accepted InstructPix2Pix [9] as our primary target model for prompt-based image editing. In our experiments, we utilize a filtered subset of the CelebA-HQ dataset [49], a high-quality human face attribute dataset widely used in the facial analysis community. The dataset consists of $2,000$ human portrait images spanning diverse race, age, and gender groups. For editing prompts, we manually selected 25 prompts across three categories: facial feature modifications (*e.g.*, hair, nose modification), accessory adjustments (*e.g.*, clothing, eyewear), and background alterations.

Baselines. We evaluate FACELOCK against two established text-guided image editing protection methods: PhotoGuard [1] and EditShield [2], both designed for general image protection. Additionally, we also compare against a variety of widely used methods [50–56] in adversarial machine learning field, including untargeted encoder attack, CW attack, and VAE attack as other baseline methods. Full details on these baselines are provided in Appx. A.

Evaluation metrics. We adopt quantitative evaluation metrics across two categories: prompt fidelity and image integrity. For prompt fidelity, we report PSNR, SSIM, and LPIPS scores between edits on protected and unprotected images, as well as the CLIP similarity score (CLIP-S), which captures the alignment between the edit-source image embedding shift and the text embedding. For image integrity, we report the CLIP image similarity score (CLIP-I) and facial recognition similarity score (FR). CLIP-I captures overall visual similarity, while FR specifically measures similarity in biometric information.

Implementation details For a fair comparison, we set the perturbation budget to 0.02 and the number of iterations to 100 for all methods, except EditShield, which does not have a default perturbation budget. Additionally, we include the untargeted latent-wise loss from EditShield as a regularization term to stabilize the protection results. Further experimental details are provided in Appx. A.

5.2. Experiment Results

Superior performance of FACELOCK in human portrait image protection: quantitative and qualitative evaluation. Building upon our analysis of comprehensive evaluation metrics for image editing and protection, we present a quantitative evaluation of various protection methods in **Tab. 1**. Our proposed method, FACELOCK, demonstrates remarkable protection effectiveness across both prompt fidelity and image integrity metrics. Regarding prompt fidelity, FACELOCK achieves competitive results in multiple metrics. It ties the lowest SSIM score and maintains a competitive CLIP-S score, and more notably, it excels in the LPIPS metric with the highest score. This aligns with our discussion on the importance of perceptual measures over pixel-based metrics. For image integrity, FACELOCK outperforms all baselines significantly, especially in FR scores. This underscores its unparalleled efficacy in protecting the subject's biometric information against malicious editing. In **Fig. 6**, we present qualitative results of the three editing types. As we can see, Our approach demonstrates the most pronounced alteration of biometric details between the edited and source images.

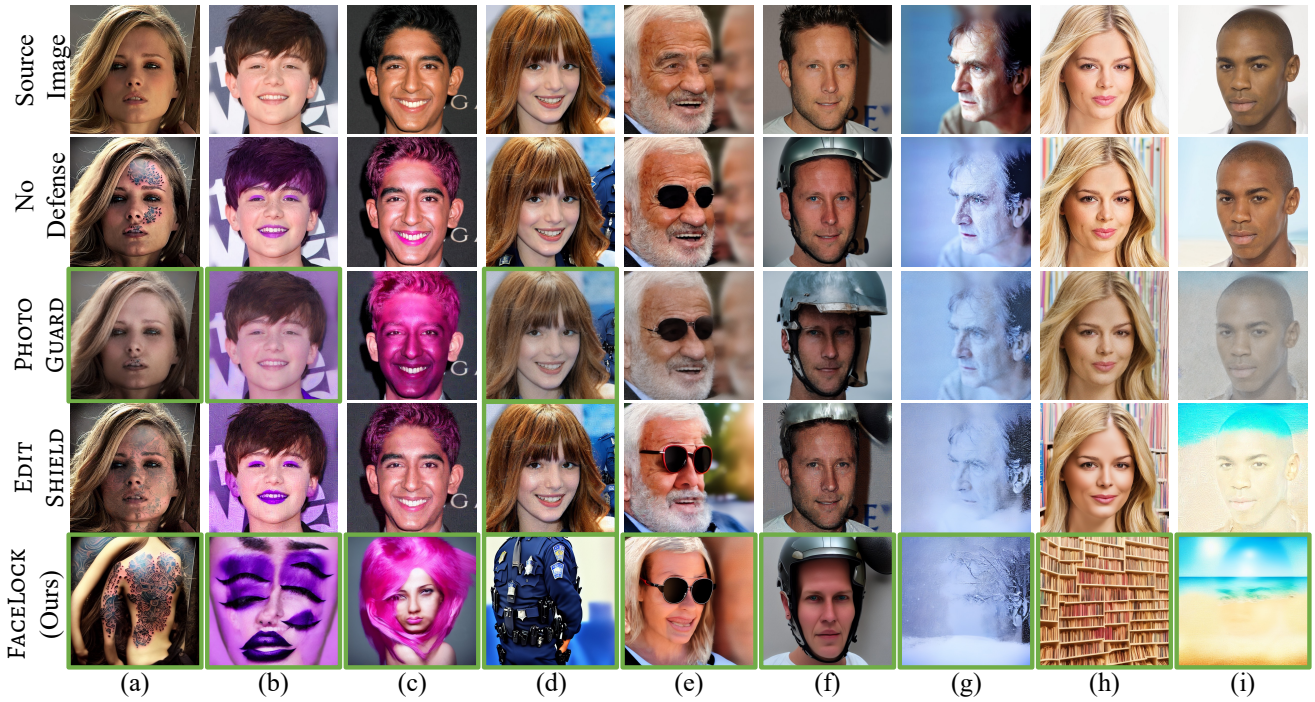

Figure 6. Qualitative results of different defense methods. Three editing types are included: **facial feature modifications** ((a) 'Let the person have a tattoo'; (b) 'Let the person wear purple makeup'; (c) 'Turn the person's hair pink'), **accessory adjustments** ((d) 'Let the person wear a police suit'; (e) 'Let the person wear sunglasses'; (f) 'Let the person wear a helmet'), and **background alterations** ((g) 'Let it be snowy'; (h) 'Set the background in a library'; (i) 'Change the background to a beach'). Images in green frames denote successful defenses.

For example, in the "*Let the person wear sunglasses*" editing scenario, while the edited image presents a person wearing sunglasses, it also transforms the individual from an elderly man in the source image to a young woman. These results indicate that FaceLock effectively protects images from different editing instructions.

FaceLock demonstrates consistent protection across diverse editing types. In **Tab.** 2, we present the facial recognition similarity scores for three editing types: facial feature modifications, accessory adjustments, and background alterations. As shown, FaceLock provides robust protection across all three editing types. Notably, background alterations generally yield the highest facial recognition similarity scores across all methods and the clean edit scenario, indicating that just modifying the background is prone to preserve more facial identity compared to direct facial modifications. Despite the inherent challenge of protecting identity during background alterations, our method still achieves promising protection results in this category, demonstrating its effectiveness even in the most demanding scenarios.

FaceLock demonstrates superior robustness against common purification techniques compared to existing methods. We examined the robustness of the defense methods, we test three commonly used heuristic purification methods: Gaussian blurring, image rotation, and JPEG compression. These techniques were applied to images with adversarial perturbations. As shown in **Tab.** 3, FaceLock consistently outperforms both PhotoGuard and EditShield across all purification techniques.

Table 2. Facial recognition similarity score (lower the better) over different editing types. Three types of editing prompts are considered, including facial feature modifications, accessory adjustments, and background alterations.

Method	Accessory	Facial Feature	Background
No Defense	0.758	0.849	0.896
PhotoGuard	0.507	0.483	0.837
EditShield	0.489	0.673	0.770
FaceLock	**0.245**	**0.339**	**0.362**

Ablation studies on perturbation budgets. To demonstrate the impact of perturbation budgets on our protection method, we conducted an ablation study by varying the budget from 0.01 to 0.05. As shown in **Tab.** 4, increasing the perturbation budget consistently reduces the facial recognition similarity score between the edit image and the source image, indicating stronger protection. However, as it is shown in **Fig.** 7, large budgets (*e.g.*, 0.05) introduce perceptible artifacts, compromising the image quality. Thus, we select a budget of 0.02 in our main experiments, which achieves effective protection with an imperceptible perturbation.

Ablation studies on the effect of different protection components. The analysis of protection components, as presented in **Tab.** 5, was conducted to evaluate the effectiveness of different elements in the perturbation optimization pro-

Table 3. Robustness comparison of image protection methods against common purification techniques. Arrows ↑ and ↓ represent a higher or lower value is preferred for a successful defense. None denotes no purification techniques applied. Blur denotes Gaussian blurring ($k = 5, \sigma = 1.5$), Rotate denotes random rotation between (-10, 10) degrees. JPEG Q denotes JPEG compression at quality level Q.

	LPIPS ↑						FR ↓					
Method	None	Blur	Rotate	JPEG 60	JPEG 75	JPEG 90	None	Blur	Rotate	JPEG 60	JPEG 75	JPEG 90
PhotoGuard	0.376	0.306	0.367	0.249	0.248	0.278	0.523	0.804	0.719	0.786	0.782	0.780
EditShield	0.370	0.295	0.356	0.231	0.285	0.318	0.585	0.763	0.663	0.744	0.713	0.645
FACELOCK	**0.439**	**0.363**	**0.405**	**0.292**	**0.302**	**0.345**	**0.308**	**0.553**	**0.544**	**0.709**	**0.624**	**0.590**

Table 4. Facial recognition similarity score FR over different perturbation budgets. A lower FR is preferred for a successful defense.

	Perturbation Budgets				
Metrics	0.01	0.02	0.03	0.04	0.05
FR ↓	0.557	0.314	0.259	0.216	0.196

Table 5. Comparison of different design configurations and their impact on LPIPS and FR metrics. Arrows ↑ and ↓ represent a higher or lower value is preferred for a successful defense.

	Components				Metrics	
Design	CVL	Diffusion	Pixel	Feature	**LPIPS ↑**	**FR ↓**
CVL	✓				0.091	0.667
CVL-D	✓	✓			0.381	0.380
CVL-DP	✓	✓	✓		0.381	0.573
FACELOCK	✓	✓		✓	**0.423**	**0.377**

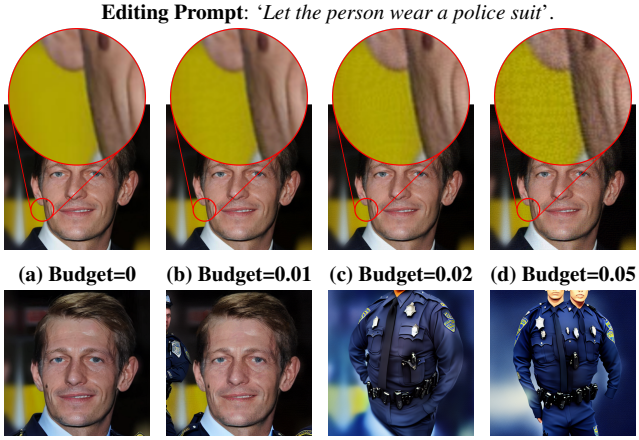

Figure 7. Protected images injected with perturbation of different budgets, along with their corresponding editing results.

cess. By examining various design configurations, we aimed to understand how each component contributes to the overall protection mechanism. The improvements in both the LPIPS metric and the FR metric from **Design I: CVL** to other design configurations showcases the importance of involving the diffusion process in the optimization loop. Interestingly, the LPIPS metric remains constant at 0.381 for both **Design II: CVL-D** and **Design III: CVL-DP** despite the addition of the pixel-level penalty in the optimization process. This observation aligns with our analysis in Sec. 3, underscoring that incorporating pixel-level loss does not disrupt the overall feature-level disparity. Furthermore, our proposed method FACELOCK achieves the best results in both LPIPS and FR metrics, indicating a better perceptual protection in both prompt fidelity and image integrity requirements.

FACELOCK's robustness to feature extractor choices. The results in **Tab. 6** highlight that FACELOCK performs consistently accross different pretrained convolutional neural networks used as feature extractors. Specifically, we observe that the LPIPS score, which measures the differences by comparing high-level semantic features of images are comparable between all three networks. Similarly, the FR scores, which assess the effectiveness of our method in disrupting biometric recognition, show that FACELOCK achieves similar performance in reducing facial recognition similarity, regardless of the network used, reinforcing that FACELOCK remains robust in its protection across different feature extractors.

Table 6. Performance comparison of FACELOCK with different pretrained CNNs used as feature extractors.

CNN	**LPIPS ↑**	**FR ↓**
AlexNet	0.451	0.346
SqueezeNet	0.451	0.345
VGG	0.458	0.340

Additional Results. We also conducted additional experiments on evaluating the generalization abilities of FACELOCK protection on different image editing tools [57, 58]. We refer more discussions in Appx. B.

6. Conclusion, Limitation, and Discussion

In this paper, we present FACELOCK, an innovative method to protect human portrait images from malicious editing by optimizing adversarial perturbations that prevent biometric recognition post-editing. FACELOCK effectively disrupts identifiable facial features, breaking the biometric link between original and edited images. Experiments show its superior performance over existing defenses and robustness against purification techniques. While FACELOCK is tailored for single portraits, extending its efficacy to images with multiple individuals remains a challenge. Additionally, emerging generative models like rectified flows [59, 60] may require further adaptations to sustain robustness. Addressing these challenges can enhance privacy protection at the forefront of generative models.

Broader Impact and Ethics Statement

Broader Impact Statement. Advancements in diffusion-based image editing enable creative expression but also pose significant risks to privacy and identity security. Our work, FACELOCK, addresses these risks by providing a robust defense mechanism that renders biometric information unrecognizable after edits. By demonstrating the potential pitfalls in current evaluation metrics, we aim to encourage the development of more reliable and effective solutions in this domain.

Ethics Statement. We believe our work sets a precedent for privacy-preserving AI research, especially in image synthesis and editing. By showing that privacy protection can be achieved through targeting biometric integrity, we hope to inspire more robust and innovative approaches to privacy in the broader context of Generative AI systems. This research also contributes to the ongoing dialogue about responsible AI development and highlights the importance of addressing privacy concerns as these technologies continue to advance.

References

[1] H. Salman, A. Khaddaj, G. Leclerc, A. Ilyas, and A. Madry, "Raising the cost of malicious ai-powered image editing," *arXiv preprint arXiv:2302.06588*, 2023. 1, 2, 3, 4, 6, 12, 14

[2] R. Chen, H. Jin, Y. Liu, J. Chen, H. Wang, and L. Sun, "Editshield: Protecting unauthorized image editing by instruction-guided diffusion models," *arXiv preprint arXiv:2311.12066*, 2023. 1, 2, 3, 4, 6, 12

[3] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in *Conference on Computer Vision and Pattern Recognition 2023*, 2023. 2

[4] Z. Zhang, L. Han, A. Ghosh, D. Metaxas, and J. Ren, "Sine: Single image editing with text-to-image diffusion models," *arXiv preprint arXiv:2212.04489*, 2022.

[5] K. Zhang, L. Mo, W. Chen, H. Sun, and Y. Su, "Magicbrush: A manually annotated dataset for instruction-guided image editing," in *Advances in Neural Information Processing Systems*, 2023.

[6] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022. 2

[7] Y. Huang, L. Xie, X. Wang, Z. Yuan, X. Cun, Y. Ge, J. Zhou, C. Dong, R. Huang, R. Zhang *et al.*, "Smartedit: Exploring complex instruction-based image editing with multimodal large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8362–8371. 2, 3

[8] T.-J. Fu, W. Hu, X. Du, W. Y. Wang, Y. Yang, and Z. Gan, "Guiding Instruction-based Image Editing via Multimodal Large Language Models," in *International Conference on Learning Representations (ICLR)*, 2024. 2

[9] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 392–18 402. 2, 3, 6, 12

[10] J. Choi, Y. Choi, Y. Kim, J. Kim, and S.-H. Yoon, "Custom-edit: Text-guided image editing with customized diffusion models," *ArXiv*, vol. abs/2305.15779, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258888143 2

[11] I. Han, S. Yang, T. Kwon, and J. C. Ye, "Highly personalized text embedding for image manipulation by stable diffusion," 2023. [Online]. Available: https://arxiv.org/abs/2303.08767 2

[12] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[13] J. Gu, Y. Wang, N. Zhao, T.-J. Fu, W. Xiong, Q. Liu, Z. Zhang, H. Zhang, J. Zhang, H. Jung, and X. E. Wang, "Photoswap: Personalized subject swapping in images," 2023. 2

[14] Z. Liu, J. Huang, H. Chu, and Q. Xu, "Swapanything: Towards human-centric face and object swapping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[15] H. Zohny, J. McMillan, and M. King, "Ethics of generative ai," pp. 79–80, 2023. 2

[16] B. Vyas, "Ethical implications of generative ai in art and the media," *International Journal for Multidisciplinary Research (IJFMR), E-ISSN*, pp. 2582–2160, 2024.

[17] G. Lawton. (2024) Generative ai ethics: 8 biggest concerns and risks. Published: 23 Jul 2024, Accessed: 2024-11-13. [Online]. Available: https://www.techtarget.com/ 2

[18] T. N. Y. Times, "Taylor swift ai fake images controversy," January 2024, accessed: 13-Nov-2024. [Online]. Available: https://www.nytimes.com/2024/01/26/arts/music/taylor-swift-ai-fake-images.html 2

[19] BBC News, "Inside the deepfake porn crisis engulfing korean schools," September 2024, accessed: 13-Nov-2024. [Online]. Available: https://www.bbc.com/news/articles/cpdlpj9zn9go 2

[20] W. Zhao, Y. Rao, W. Shi, Z. Liu, J. Zhou, and J. Lu, "Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8568–8577. 2

[21] L. Tang, W. Ma, M. Grobler, W. Meng, Y. Wang, and S. Wen, "Faces are protected as privacy: An automatic tagging framework against unpermitted photo sharing in social media," *IEEE Access*, vol. 7, pp. 75 556–75 567, 2019. 2

[22] J. An, W. Zhang, D. Wu, Z. Lin, J. Gu, and W. Wang, "Sd4privacy: exploiting stable diffusion for protecting facial privacy," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[23] X. He, M. Zhu, D. Chen, N. Wang, and X. Gao, "Diff-privacy: Diffusion-based face privacy protection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2

[24] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Glaze: Protecting artists from style mimicry by text-to-image models," *USENIX Security Symposium*, 2023. 2, 3, 4

[25] R. Wang, H. Chang, D. Gandikota, and S. Jha, "Distraction is all you need: Instruction-based image editing with complementary attention," *arXiv preprint arXiv:2306.05934*, 2023. 3

[26] E. Huang and B. Y. Zhao, "Nightshade: A data poisoning tool to protect artists from generative ai," *arXiv preprint arXiv:2310.13828*, 2023. 3, 4

[27] R. Wang, P. Ghosh, and S. Jha, "Advdm: Generating adversarial examples for diffusion models," *arXiv preprint arXiv:2305.16317*, 2023. 2, 3, 4

[28] J. Zhang, Z. Xu, S. Cui, C. Meng, W. Wu, and M. R. Lyu, "On the robustness of latent diffusion models," *arXiv preprint arXiv:2306.08257*, 2023. 2

[29] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695. 2, 3

[30] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013, vol. 87. 3, 4

[31] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy (SP)*, 2017. 3

[32] M. Kim, A. K. Jain, and X. Yu, "Adaface: Quality adaptive margin for face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3, 4

[33] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[34] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[35] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "Elasticface: Elastic margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2022, pp. 1578–1587.

[36] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, and F. H. Jilin Li, "Curricularface: Adaptive curriculum learning loss for deep face recognition," pp. 1–8, 2020.

[37] P. Terhörst, M. Ihlefeld, M. Huber, N. Damer, F. Kirchbuchner, K. Raja, and A. Kuijper, "QMagFace: Simple and accurate quality-aware face recognition," *CoRR*, vol. abs/2111.13475, 2021. [Online]. Available: https://arxiv.org/abs/2111.13475

[38] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "MagFace: A universal representation for face recognition and quality assessment," in *CVPR*, 2021. 3

[39] F. Boutros, M. Huber, P. Siebke, T. Rieber, and N. Damer, "Sface: Privacy-friendly and accurate face recognition using synthetic data," in *IEEE International Joint Conference on Biometrics, IJCB 2022, Abu Dhabi, United Arab Emirates, October 10-13, 2022*. IEEE, 2022, pp. 1–11. [Online]. Available: https://doi.org/10.1109/IJCB54206.2022.10007961 3

[40] F. Boutros, M. Huber, A. T. Luu, P. Siebke, and N. Damer, "Sface2: Synthetic-based face recognition with w-space identity-driven sampling," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, pp. 1–1, 2024.

[41] F. Boutros, M. Klemt, M. Fang, A. Kuijper, and N. Damer, "Unsupervised face recognition using unlabeled synthetic data," in *17th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2023, Waikoloa Beach, HI, USA, January 5-8, 2023*. IEEE, 2023, pp. 1–8. [Online]. Available: https://doi.org/10.1109/FG57933.2023.10042627

[42] J. N. Kolf, T. Rieber, J. Elliesen, F. Boutros, A. Kuijper, and N. Damer, " Identity-driven Three-Player Generative Adversarial Network for Synthetic-based Face Recognition ," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2023, pp. 806–816. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPRW59228.2023.00088

[43] H. Qiu, B. Yu, D. Gong, Z. Li, W. Liu, and D. Tao, "Synface: Face recognition with synthetic data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 880–10 890. 3

[44] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar, "Diffusion models for adversarial purification," *arXiv preprint arXiv:2205.07460*, 2022. 4

[45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004. 5

[46] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763. 5

[47] Y. Zhang, Y. Zhang, Y. Yao, J. Jia, J. Liu, X. Liu, and S. Liu, "Unlearncanvas: A stylized image dataset to benchmark machine unlearning for diffusion models," *arXiv preprint arXiv:2402.11846*, 2024. 5

[48] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595. 5

[49] T. Karras, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017. 6

[50] Y. Zhang, G. Zhang, P. Khanduri, M. Hong, S. Chang, and S. Liu, "Revisiting and advancing fast adversarial training through the lens of bi-level optimization," in *International Conference on Machine Learning*. PMLR, 2022, pp. 26 693–26 712. 6

[51] G. Zhang, Y. Zhang, Y. Zhang, W. Fan, Q. Li, S. Liu, and S. Chang, "Fairness reprogramming," *Advances in Neural*

Information Processing Systems, vol. 35, pp. 34 347–34 362, 2022.

[52] G. Zhang, S. Lu, Y. Zhang, X. Chen, P.-Y. Chen, Q. Fan, L. Martie, L. Horesh, M. Hong, and S. Liu, "Distributed adversarial training to robustify deep neural networks at scale," in *Uncertainty in artificial intelligence*. PMLR, 2022, pp. 2353–2363.

[53] Y. Zhang, R. Cai, T. Chen, G. Zhang, H. Zhang, P.-Y. Chen, S. Chang, Z. Wang, and S. Liu, "Robust mixture-of-expert training for convolutional neural networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 90–101.

[54] Y. Zhang, J. Jia, X. Chen, A. Chen, Y. Zhang, J. Liu, K. Ding, and S. Liu, "To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now," *arXiv preprint arXiv:2310.11868*, 2023.

[55] Y. Zhang, X. Chen, J. Jia, Y. Zhang, C. Fan, J. Liu, M. Hong, K. Ding, and S. Liu, "Defensive unlearning with adversarial training for robust concept erasure in diffusion models," *arXiv preprint arXiv:2405.15234*, 2024.

[56] H. Zhuang, Y. Zhang, and S. Liu, "A pilot study of query-free adversarial attack against stable diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2385–2392. 6

[57] J. Gu, Y. Wang, N. Zhao, W. Xiong, Q. Liu, Z. Zhang, H. Zhang, J. Zhang, H. Jung, and X. E. Wang, "Swapanything: Enabling arbitrary object swapping in personalized visual editing," *arXiv preprint arXiv:2404.05717*, 2024. 8

[58] T.-J. Fu, W. Hu, X. Du, W. Y. Wang, Y. Yang, and Z. Gan, "Guiding instruction-based image editing via multimodal large language models," *arXiv preprint arXiv:2309.17102*, 2023. 8

[59] X. Liu, C. Gong, and Q. Liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," *arXiv preprint arXiv:2209.03003*, 2022. 8

[60] L. Rout, Y. Chen, N. Ruiz, C. Caramanis, S. Shakkottai, and W.-S. Chu, "Semantic image inversion and editing using rectified stochastic differential equations," *arXiv preprint arXiv:2410.10792*, 2024. 8

Appendix

A. Detailed Experiment Setups

A.1. Implementation Details of FACELOCK

FACELOCK optimizes perturbation on facial disruption and feature embedding disparity that prevent biometric recognition post-editing. The pseudocode of FACELOCK is presented in Algorithm 1. More specifically, the facial recognition loss function f_{FR} is defined as the negative of the similarity score between the input images computed by the CVLFACE model[1], and the feature disparity loss function f_{FE} is computed as the weighted sum of the layer-wise feature embedding distances across the feature extractor network. As mentioned in Sec 5, we also include the untargeted latent-wise loss from EditShield[2] as a regularization term to stabilize the protection results. The hyper-parameters used in our implementation are summarized in **Tab A1**.

Algorithm 1 FACELOCK

Input: Input image $\mathbf{x}$, VAE $\mathcal{E}, \mathcal{D}$ in the diffusion model, step size α, number of steps N, overall perturbation budget ϵ, regularization weight λ, facial recognition loss function f_{FR}, feature disparity loss function f_{FE}

Initialize perturbation $\boldsymbol{\delta} \leftarrow N(0, \mathbf{I})$, and the protected image $\mathbf{x}' \leftarrow \mathbf{x} + \boldsymbol{\delta}$
Compute the latent embedding of the input image $\mathbf{z} \leftarrow \mathcal{E}(\mathbf{x})$
for $n = 1$ to N **do**
    Compute the latent embedding of the protected image $\mathbf{z}' \leftarrow \mathcal{E}(\mathbf{x}')$
    Compute the decoded image from the latent embedding $\mathbf{x}_d \leftarrow \mathcal{D}(\mathbf{z}')$
    Compute the facial recognition loss $l_{\text{FR}} \leftarrow f_{\text{FR}}(\mathbf{x}_d, \mathbf{x})$
    Compute the feature disparity loss $l_{\text{FE}} \leftarrow f_{\text{FE}}(\mathbf{x}_d, \mathbf{x})$
    Compute the latent loss (regularization term) $l_{\text{L}} \leftarrow \|\mathbf{z}' - \mathbf{z}\|_2^2$
    Update the perturbation $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta} + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}'}(l_{\text{FR}} + l_{\text{FE}} + \lambda \cdot l_{\text{L}}))$
    $\boldsymbol{\delta} \leftarrow \text{clip}(\boldsymbol{\delta}, -\epsilon, \epsilon)$
    Update the protected image: $\mathbf{x}' \leftarrow \mathbf{x} + \boldsymbol{\delta}$
end for

Return: The protected image $\mathbf{x}'$

Table A1. Hyper-parameters used for the implementation.

Norm	**perturbation budget** ϵ	**step size** α	**number of steps** N	λ
l_∞	0.02	0.003	100	0.2

A.2. Implementation Details of Baselines

In addition to using previous methods [1, 2] as baselines, we also compare our FACELOCK approach against several widely used techniques in the adversarial machine learning field. These methods are summarized in Algorithms 2, 3, and 4. To ensure a fair comparison, we use the same hyper-parameters settings in Tab A1.

A.3. Image Editing Details

Models. For image editing, we use the open-source instruction-guided diffusion model InstructPix2Pix [9] hosted on Hugging Face[2] as our primary target model. We use the hyper-parameters presented in **Tab A2**. We use the same seed setting when comparing edits on the unprotected images and the images protected by different methods to ensure that the edit images are are modified in the same way and that the different editing effects are due to the protection methods instead of random seeds.
Dataset. For the human portrait images used in our experiments, we utilize a filtered subset of the CelebA-HQ dataset[3], a high-quality human face attribute dataset widely used in the facial analysis community. The dataset consists of $2,000$ human portrait images ensuring diversity across various demographic groups, including race, age, and gender, to enhance

[1]The model is available on https://github.com/mk-minchul/CVLface
[2]The model is available on https://huggingface.co/timbrooks/instruct-pix2pix
[3]The dataset is available on https://www.kaggle.com/datasets/lamsimon/celebahq/data

Algorithm 2 Untargeted Encoder Attack

Input: Input image $\mathbf{x}$, VAE $\mathcal{E}$ in the diffusion model, step size α, number of steps N, overall perturbation budget ϵ

Initialize perturbation $\boldsymbol{\delta} \leftarrow N(0, \mathbf{I})$, and the protected image $\mathbf{x}' \leftarrow \mathbf{x} + \boldsymbol{\delta}$
Compute the latent embedding of the input image $\mathbf{z} \leftarrow \mathcal{E}(\mathbf{x})$
for $n = 1$ to N **do**
    Compute the latent embedding of the protected image $\mathbf{z}' \leftarrow \mathcal{E}(\mathbf{x}')$
    Compute the latent loss $l \leftarrow \|\mathbf{z}' - \mathbf{z}\|_2^2$
    Update the perturbation $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta} + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}'} l)$
    $\boldsymbol{\delta} \leftarrow \text{clip}(\boldsymbol{\delta}, -\epsilon, \epsilon)$
    Update the protected image $\mathbf{x}' \leftarrow \mathbf{x} + \boldsymbol{\delta}$
end for

Return: The protected image $\mathbf{x}'$

Algorithm 3 VAE Attack

Input: Input image $\mathbf{x}$, target image $\mathbf{x}_{\text{tgt}}$,VAE $\mathcal{E}$, $\mathcal{D}$ in the diffusion model, step size α, number of steps N, overall perturbation budget ϵ

Initialize perturbation $\boldsymbol{\delta} \leftarrow N(0, \mathbf{I})$, and the protected image $\mathbf{x}' \leftarrow \mathbf{x} + \boldsymbol{\delta}$
for $n = 1$ to N **do**
    Compute the decoded image $\mathbf{x}_d \leftarrow \mathcal{D}(\mathcal{E}(\mathbf{x}'))$
    Compute the loss $l \leftarrow \|\mathbf{x}_d - \mathbf{x}_{\text{tgt}}\|_2^2$
    Update the perturbation $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta} - \alpha \cdot \text{sign}(\nabla_{\mathbf{x}'} l)$
    $\boldsymbol{\delta} \leftarrow \text{clip}(\boldsymbol{\delta}, -\epsilon, \epsilon)$
    Update the protected image $\mathbf{x}' \leftarrow \mathbf{x} + \boldsymbol{\delta}$
end for

Return: The protected image $\mathbf{x}'$

Algorithm 4 CW L_2 Attack

Input: Input image $\mathbf{x}$, VAE $\mathcal{E}$ in the diffusion model, step size α, number of steps N, overall perturbation budget ϵ, weight c

Initialize $\mathbf{w} \leftarrow \mathbf{0}$
Compute the latent embedding of the input image: $\mathbf{z} \leftarrow \mathcal{E}(\mathbf{x})$
for $n = 1$ to N **do**
    Compute the protected image $\mathbf{x}' \leftarrow \frac{1}{2}(\tanh(\mathbf{w}) + 1)$
    Compute the latent embedding of the protected image $\mathbf{z}' \leftarrow \mathcal{E}(\mathbf{x}')$
    Compute the L_2 loss $l_{L_2} \leftarrow \|\mathbf{x}' - \mathbf{x}\|_2^2$
    Compute the latent loss $l_{\text{L}} \leftarrow -\|\mathbf{z}' - \mathbf{z}\|_2^2$
    Update $\mathbf{w} \leftarrow \mathbf{w} - \alpha \cdot \nabla_{\mathbf{w}}(l_{L_2} + c \cdot l_L)$
end for
Compute $\boldsymbol{\delta} \leftarrow \text{clip}(\frac{1}{2}(\tanh(\mathbf{w}) + 1) - \mathbf{x}, -\epsilon, \epsilon)$
Compute the protected image $\mathbf{x}' \leftarrow \mathbf{x} + \boldsymbol{\delta}$

Return: The protected image $\mathbf{x}'$

the representativeness of our experiments. For the editing prompts, we manually selected 25 prompts across three categories: facial feature modification, accessory adjustments, and background alternations. These prompts were specifically selected to produce noticeable changes across a wide range of images, avoiding those that would fail to affect a certain subset (*e.g.*, "*Let the person wear glasses*" will be ineffective for individuals who already wear glasses, which is a significant portion of the dataset). The specific prompts utilized in our experiments are listed in **Table A3** for detailed reference.

Table A2. Hyper-parameters used for the image editing process.

image size	inference steps	image guidance scale	text guidance scale
512×512	50	1.5	7.5

Table A3. Editing prompts categorized into facial feature modifications, accessory adjustments, and background alterations.

Category	Prompts
Facial Feature Modifications	❶ Turn the person's hair pink; ❷ Let the person turn bald; ❸ Let the person have a tattoo; ❹ Let the person wear purple makeup; ❺ Let the person grow a mustache; ❻ Turn the person into a zombie; ❼ Change the skin color to Avatar blue; ❽ Add elf-like ears; ❾ Add large vampire fangs; ❿ Apply Goth style makeup.
Accessory Adjustments	❶ Let the person wear a police suit; ❷ Let the person wear a bowtie; ❸ Let the person wear a helmet; ❹ Let the person wear sunglasses; ❺ Let the person wear earrings; ❻ Let the person smoke a cigar; ❼ Place a headband in the hair; ❽ Place a tiara on the top of the head.
Background Alterations	❶ Let it be snowy; ❷ Change the background to a beach; ❸ Add a city skyline background; ❹ Add a forest background; ❺ Change the background to a desert; ❻ Set the background in a library; ❼ Let the person stand under the moon;

A.4. Evaluation Metrics

PSNR, SSIM, and LPIPS scores. In our experiments, we compute the PSNR and SSIM scores using the torchmetrics library[4], while the LPIPS score is computed using the lpips library[5]. All these three metrics are computed by comparing the similarity between the edited image without defense and the edited image with defense. A lower similarity score (lower PSNR, SSIM score and higher LPIPS score) indicates better protection. PSNR and SSIM primarily focus on pixel-level statistical information, while LPIPS evaluates the similarity of high-level semantic features, capturing perceptual differences that are more aligned with human visual perception.

CLIP-S score. In the main paper, we utilize the CLIP-S metric to assess the prompt fidelity by computing the similarity between the image embedding shift and the text embedding in the CLIP embedding space:

$$\text{CLIP-S} = \frac{(E_{\text{edit}} - E_{\text{src}}) \cdot E_{\text{prompt}}}{\|E_{\text{edit}} - E_{\text{src}}\| \|E_{\text{prompt}}\|}, \tag{A1}$$

where E_{src} denotes the CLIP image embedding of the source image, E_{edit} denotes the CLIP image embedding of the edited image, and E_{prompt} denotes the CLIP text embedding of the prompt instruction. This formulation is particularly suitable for our experiments because the prompts are designed as instructions describing the expected transformation or modification from the source image to the edited image.

CLIP-SD score. Following PhotoGuard's evaluation metric [1], an alternative approach to assess the prompt fidelity is to compute the cosine similarity directly between the embedding of the edited image and the embedding of the descriptive text prompt in the CLIP embedding space:

$$\text{CLIP-SD} = \frac{E_{\text{edit}} \cdot E_{\text{desc}}}{\|E_{\text{edit}}\| \|E_{\text{desc}}\|}, \tag{A2}$$

where E_{desc} denotes the CLIP text embedding of the descriptive text prompt. We report the CLIP-SD score for each method in **Tab A4**. From the table, we observe that, except for the VAE method, all defense methods show a worse defense effect compared to the "No Defense" scenario. This aligns with the analysis presented in Sec 4, where we discussed how CLIP-based similarity metrics often overemphasize the elements from the prompt, leading to a prioritization of over-editing. To generate the descriptive text prompts, we leverage ChatGPT based on the prompt instructions provided in Tab A3.

Table A4. Quantitative evaluation on prompt fidelity using CLIP-SD. The ↓ indicates that a lower CLIP-SD score is preferred for a successful defense.

Method	No Defense	PhotoGuard	EditShield	Untargeted Encoder	CW L2	VAE	FACELOCK(ours)
CLIP-SD↓	0.272±0.029	0.283±0.029	0.277±0.027	0.284±0.024	0.277±0.027	0.270±0.029	0.283±0.024

[4]This library can be installed from https://lightning.ai/docs/torchmetrics/stable/
[5]This library can be installed from https://pypi.org/project/lpips/

CLIP-I score. In the main paper, we utilize the CLIP-I metric to assess the image integrity by computing the similarity between the edited image embedding and the source image embedding in the CLIP embedding space:

$$\text{CLIP-I} = \frac{E_{\text{edit}} \cdot E_{\text{src}}}{\|E_{\text{edit}}\| \|E_{\text{src}}\|}. \tag{A3}$$

The CLIP-I metric is used as a general indicator of the preservation effect, providing an overall measure of how similar the edited image is to the source image in the CLIP embedding space. While this serves as a useful first step in generally evaluating image integrity, it does not specifically address biometric integrity, which is central to protecting human portrait images.

FR score. In the main paper, we utilize the CVLFACE model to compute the facial recognition similarity score between the edited and source image to indicate the preservation effect of biometric integrity:

$$\text{FR} = \text{CVLFACE}(I_{\text{edit}}, I_{\text{src}}), \tag{A4}$$

where I_{src} denotes the source image, and I_{edit} denotes the edited image. Unlike other general image similarity metrics, the CVLFACE model is tailored to assess the consistency of facial features, making it more suitable for evaluating how well the identity of the person is preserved after the image has been edited. The FR score plays a key role in assessing whether the protection method effectively disrupts the biometric identity of the person in the image.

B. Additional Experiment Results

B.1. Qualitative Results on Background Alternation

Figure A1. Qualitative results of background alternation edits across various defense methods. Images in green frames denote successful defense.

B.2. Qualitative Results on Accessory Adjustment

Figure A2. Qualitative results of accessory adjustment edits across various defense methods. Images in green frames denote successful defense.

B.3. Qualitative Results on Facial Feature Modification

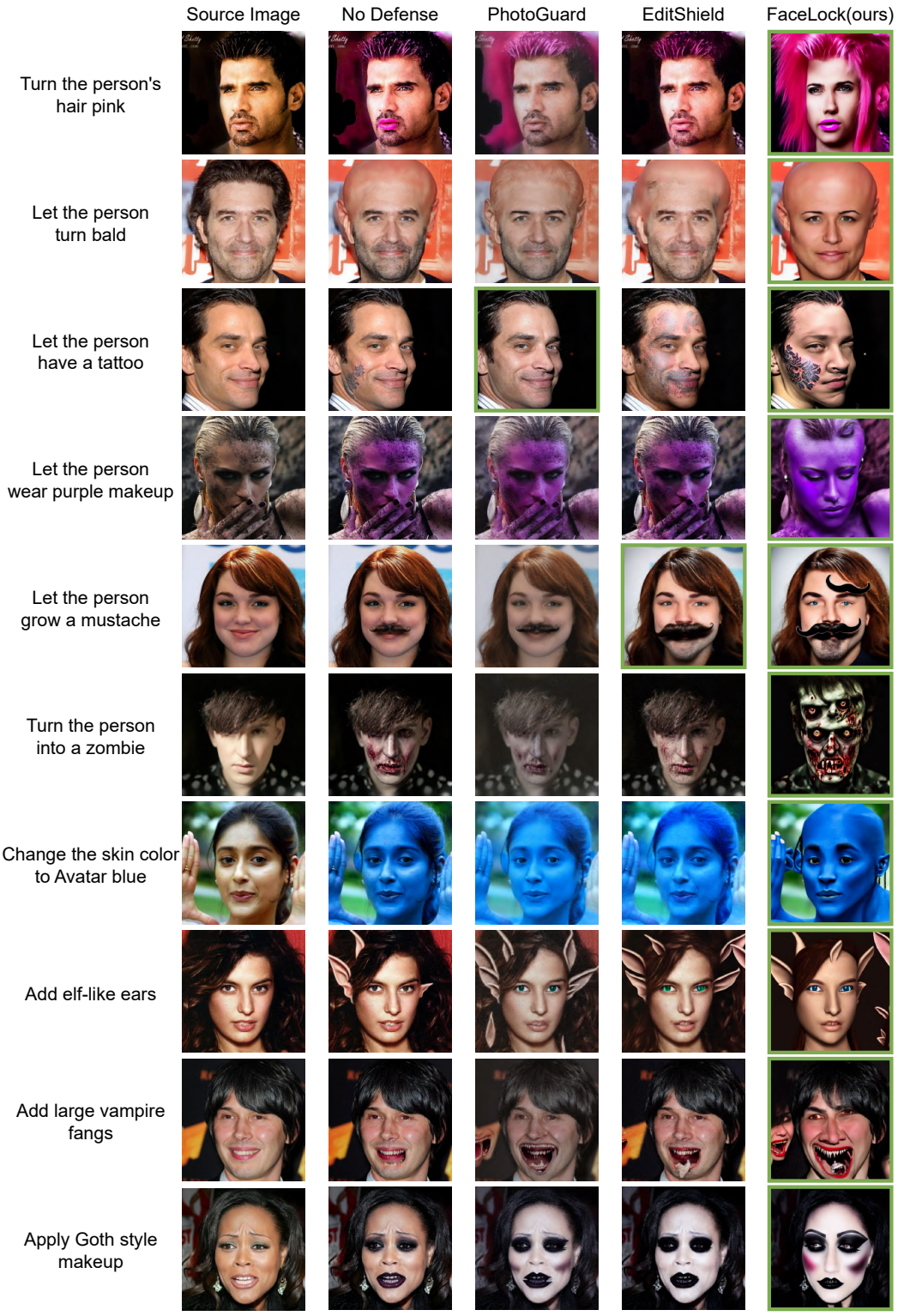

Figure A3. Qualitative results of facial feature modification edits across various defense methods. Images in green frames denote successful defense.

B.4. Qualitative Results Against Purification

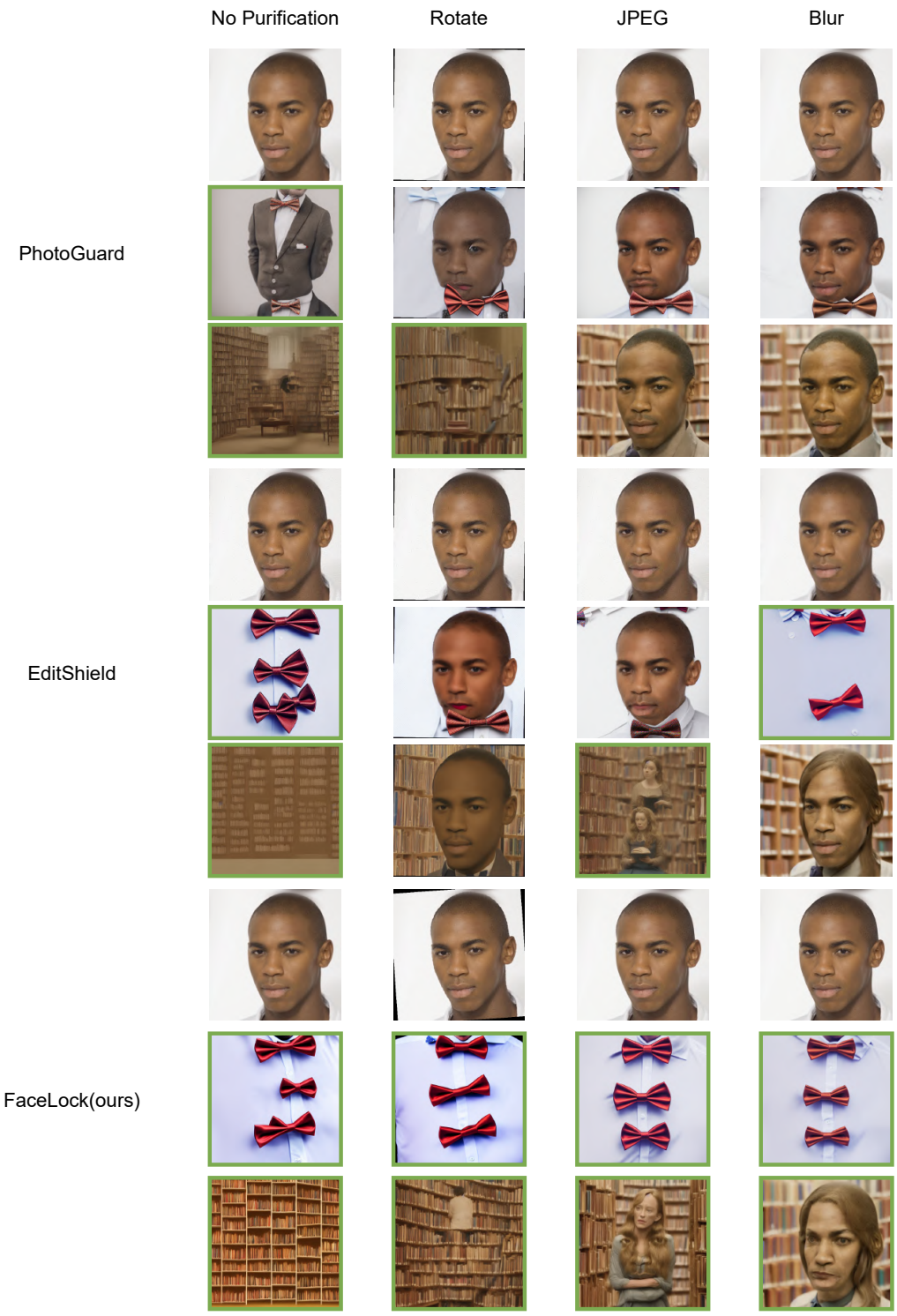

Figure A4. Qualitative results of edits on protected images after applying purification methods. Each block shows: purified protected images (1st row), edits with the instruction "Let the person wear a bowtie", and edits with the instruction "Set the background in a library". Purification methods include random rotation (-10, 10), JPEG compression (quality 75), and Gaussian blurring ($k = 5, \sigma = 1.5$). Images in green frames denote successful defense.